Lec 19 Singel cell & Visualization

Single-cell data analytics

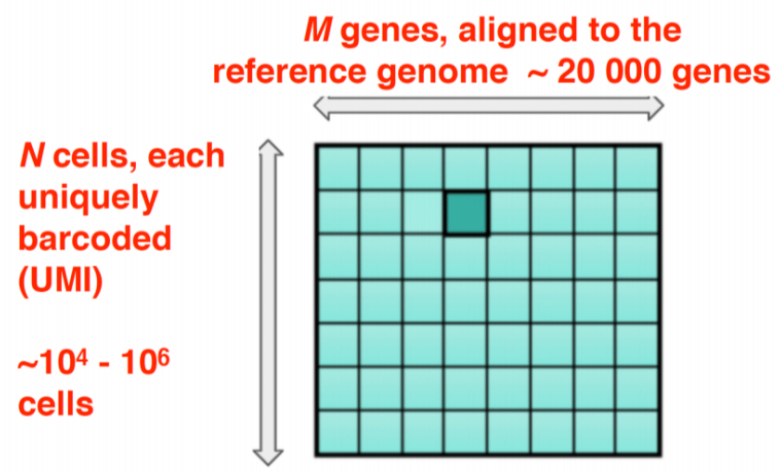
Quality control → Normalization → Feature selection → Dimension reduction → Clustering

This is a process of visualization.

<u>Challenges in single cell data analytics:</u>

1. Noise
   It may come from the mitochondrial genes or other source.
2. Doublet
   During the cell separation process, sometimes two cells can't be separated successfully, which causes the doublet. The doublet may be misleading because researchers may see doublet as a new cluster.
   Methods of removing doublet: identify doublet data points using the simulated doublet data points.
3. Dropout (different from dropout in machine learning)
   This means that in the sequencing process, the data of some genes is empty. This may result from a limited budget. To address this issue, one method is seeing the dropout as missing value and using statistical or machine learning methods to complement it.
4. Batch effect
   This may be caused by the some non-biological factors. For example, one laboratory has more budget so it can reach higher sequence depth, while another laboratory does not have much budget so its sequence depth is lower.
   There are some approaches like normalization and alignment that may be helpful for this problem.

Gene expression matrix:



CPM(counts per million) = $10^6 * X_i / N$

The reason that it needs to time a 1 million is that counts in each cell is too small.

Visualization

One method of dimension reduction is PCA, but it has a problem. That is it can't preserve the original cluster information. To preserve the cluster information, t-SNE is used.

T-SNE

Its key is that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.

The process of t-SNE

1. Random initialization
2. For each point, update the position a little bit
3. Compare the cluster to the original cluster. The points from the same cluster attract each other. The points from different clusters push apart each other
4. Until no more update

However, there are also some disadvantages of t-SNE. For example, longer running time and not preserved original distances.

Protein-RNA/DNA interaction

The way to get the binding motif (means the preference of protein binding): bind → pull-down → sequence

From aligned sequences to motif:

1. Using multiple alignment to align sequences

Table 1: Starting sequences.

#	Sequence
1	AAGAAT
2	ATCATA
3	AAGTAA
4	AACAAA
5	ATTAAA
6	AAGAAT

2. Convert to position count matrix

Table 2: Position Count Matrix.

Position	1	2	3	4	5	6
A	6	4	0	5	5	4
C	0	0	2	0	0	0
G	0	0	3	0	0	0
T	0	2	1	1	1	2

3. Convert to position probability matrix

Table 3: Position Probability Matrix.

Position	1	2	3	4	5	6
A	1.00	0.67	0.00	0.83	0.83	0.66
C	0.00	0.00	0.33	0.00	0.00	0.00
G	0.00	0.00	0.50	0.00	0.00	0.00
T	0.00	0.33	0.17	0.17	0.17	0.33

4. Draw the motif

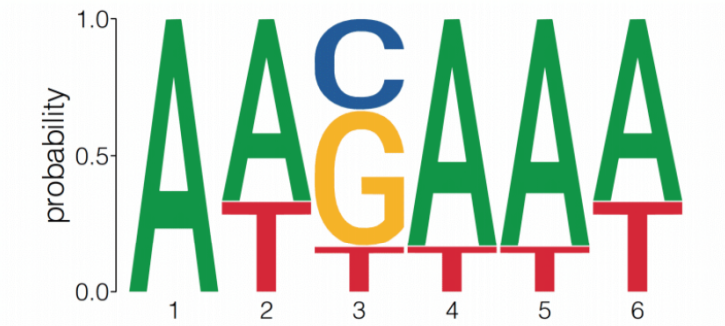


Figure 1: Sequence logo of a Position Probability Matrix